

Research Article

Talking-Face Identity Verification, Audiovisual Forgery, and Robustness Issues

Walid Karam,¹ Hervé Bredin,² Hanna Greige,³ Gérard Chollet,⁴ and Chafic Mokbel¹

¹ Computer Science Department, University of Balamand, 100 El-Koura, Lebanon

² SAMoVA Team, IRIT-UMR 5505, CNRS, 5505 Toulouse, France

³ Mathematics Department, University of Balamand, 100 El-Koura, Lebanon

⁴ TSI, Ecole Nationale Supérieure des Télécommunications, 46 rue Barrault, 75634 Paris, France

Correspondence should be addressed to Walid Karam, walid@balamand.edu.lb

Received 1 October 2008; Accepted 3 April 2009

Recommended by Kevin Bowyer

The robustness of a biometric identity verification (IV) system is best evaluated by monitoring its behavior under impostor attacks. Such attacks may include the transformation of one, many, or all of the biometric modalities. In this paper, we present the transformation of both speech and visual appearance of a speaker and evaluate its effects on the IV system. We propose *MixTrans*, a novel method for voice transformation. *MixTrans* is a mixture-structured bias voice transformation technique in the cepstral domain, which allows a transformed audio signal to be estimated and reconstructed in the temporal domain. We also propose a face transformation technique that allows a frontal face image of a client speaker to be animated. This technique employs principal warps to deform defined MPEG-4 facial feature points based on determined facial animation parameters (FAPs). The robustness of the IV system is evaluated under these attacks.

Copyright © 2009 Walid Karam et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

With the emergence of smart phones and third and fourth generation mobile and communication devices, and the appearance of a “first generation” type of mobile PC/PDA/phones with biometric identity verification, there has been recently a greater attention to secure communication and to guarantee the robustness of embedded multimodal biometric systems. The robustness of such systems promises the viability of newer technologies that involve e-voice signatures, e-contracts that have legal values, and secure and trusted data transfer regardless of the underlying communication protocol. Realizing such technologies require reliable and error-free biometric identity verification systems.

Biometric identity verification (IV) systems are starting to appear on the market in various commercial applications. However, these systems are still operating with a certain measurable error rate that prevents them from being used in a full automatic mode and still require human intervention and further authentication. This is primarily due to the

variability of the biometric traits of humans over time because of growth, aging, injury, appearance, physical state, and so forth. Impostors attempting to be authenticated by an IV system to gain access to privileged resources could take advantage of the non-zero error rate of the system by imitating, as closely as possible, the biometric features of a genuine client.

The purpose of this paper is threefold. (1) It evaluates the performance of IV systems by monitoring their behavior under impostor attacks. Such attacks may include the transformation of one, many, or all of the biometric modalities, such as face or voice. This paper provides a brief review of IV techniques and corresponding evaluations and focuses on a statistical approach (GMM). (2) It also introduces *MixTrans*, a novel mixture-structure bias voice transformation technique in the cepstral domain, which allows a transformed audio signal to be estimated and reconstructed in the temporal domain. (3) It proposes a face transformation technique that allows a 2D face image of the client to be animated. This technique employs principal warps to deform defined MPEG-4 facial feature points based

on determined facial animation parameters (FAPs). The BANCA database is used to test the effects of voice and face transformation on the IV system.

The rest of the paper is organized as follows. Section 2 introduces the performance evaluation, protocols, and the BANCA database. Section 3 is a discussion of audiovisual identity verification techniques based on Gaussian Mixture Models. Section 4 describes the imposture techniques used, including MixTrans, a novel voice transformation technique, and face transformation based on an MPEG-4 face animation with thin-plate spline warping. Section 5 discusses the experimental results on the BANCA audiovisual database. Section 6 wraps up with a conclusion.

2. Evaluation Protocols

Evaluation of audiovisual IV systems and the comparison of their performances require the creation of a reproducible evaluation framework. Several experimental databases have been set up for this purpose. These databases consist of a large collection of biometric samples in different scenarios and quality conditions. Such databases include BANCA [1], XM2VTS [2], BT-DAVID [3], BIOMET [4], and PDDatabase [5].

2.1. The BANCA Database. In this work, audiovisual verification experiments and imposture were primarily conducted on the BANCA Database [1]. BANCA is designed for testing multimodal identity verification systems. It consists of video and speech data for 52 subjects (26 males, 26 females) in four different European languages (English, French, Italian, and Spanish). Each language set and gender was divided into two independent groups of 13 subjects (denoted g1 and g2). Each subject recorded a total of 12 sessions, for a total of 208 recordings. Each session contains two recordings: a true client access and an informed impostor attack (the client proclaims in his own words to be someone else). Each subject was prompted to say 12 random number digits, his or her name, address, and date of birth.

The 12 sessions are divided into three different scenarios.

- (i) *Scenario c (controlled)*. Uniform blue background behind the subject with a quiet environment (no background noise). The camera and microphone used are of good quality (sessions 1–4).
- (ii) *Scenario d (degraded)*. Low quality camera and microphone in an “adverse” environment (sessions 5–8).

- (iii) *Scenario a (adverse)*. Cafeteria-like atmosphere with activities in the background (people walking or talking behind the subject). The camera and microphone used are also of good quality (sessions 9–12).

BANCA has also a world model of 30 other subjects, 15 males and 15 females.

Figure 1 shows example images from the English database for two subjects in all three scenarios.

The BANCA evaluation protocol defines seven distinct training/test configurations, depending on the actual conditions corresponding to training and testing. These experimental configurations are Matched Controlled (MC), Matched Degraded (MD), Matched Adverse (MA), Unmatched Degraded (UD), Unmatched Adverse (UA), Pooled Test (P), and Grand Test (G) (Table 1).

The results reported in this work reflect experiments on the “Pooled test,” also known as the “P” protocol, which is BANCA’s most “difficult” evaluation protocol: world and client models are trained on session 1 only (controlled environment), while tests are performed in all different environments (Table 1).

2.2. Performance Evaluation. The evaluation of a biometric system performance and its robustness to imposture is measured by the rate of errors it makes during the recognition process. Typically, a recognition system is a “comparator” that compares the biometric features of a user with a given biometric reference and gives a “score of likelihood.” A decision is then taken based on that score and an adjustable defined acceptance “threshold.” Two types of error rates are traditionally used.

- (i) *False Acceptance Rate (FAR)*. The FAR is the frequency that an impostor is accepted as a genuine client. The FAR for a certain enrolled person n is measured as

$$\begin{aligned} \text{FAR}(n) &= \frac{\text{Number of successful haux attempts against a person } n}{\text{Number of all haux attempts against a person } n}, \end{aligned} \quad (1)$$

and for a population of N persons, $\text{FAR} = (1/N) \sum_{n=1}^N \text{FAR}(n)$.

- (ii) *False Rejection Rate (FRR)*. The FRR is the frequency that a genuine client is rejected as an impostor:

$$\begin{aligned} \text{FRR}(n) &= \frac{\text{Number of rejected verification attempts a genuine person } n}{\text{Number of all verification attempts a genuine person } n}, \\ \text{FRR} &= \frac{1}{N} \sum_{n=1}^N \text{FRR}(n). \end{aligned} \quad (2)$$

TABLE 1: Summary of the 7 training/testing configurations of BANCA.

	Test Sessions	Train Sessions			
		1	5	9	1, 5, 9
Client	2–4	MC			
Impostor	1–4				
Client	6–8	UD	MD		
Impostor	5–8				
Client	10–12	UA		MA	
Impostor	9–12				
Client	2–4, 6–8, 10–12	P			G
Impostor	1–12				

To assess visually the performance of the authentication system, several curves are used: the Receiver Operating Characteristic (ROC) curve [6, 7], the Expected Performance Curve (EPC) [8], and the Detection error trade-off (DET) curve [9]. The ROC curve plots the *sensitivity* (fraction of true positives) of the binary classifier system versus *specificity* (fraction of false positives) as a function of the threshold. The closer the curve to 1 is, the better the performance of the system is.

While ROC curves use a biased measure of performance (EER), the EPC introduced in [8] provides an unbiased estimate of performance at various operating points.

The DET curve is a log-deviate scale graph of *FRR* versus *FAR* as the threshold changes. The *EER* value is normally reported on the DET curve: the closer *EER* to the origin is, the better the performance of the system is. The results reported in this work are in the form of DET curves.

3. Multimodal Identity Verification

3.1. Identification Versus Verification. Identity recognition can be divided into two major areas: authentication and Identification. Authentication, also referred to as verification, attempts to verify a person's identity based on a claim. On the other hand, identification attempts to find the identity of an unknown person in a set of a number of persons. Verification can be thought of as being a one-to-one match where the person's biometric traits are matched against one template (or a template of a general "world model") whereas identification is a one-to-many match process where biometric traits are matched against many templates.

Identity verification is normally the target of applications that entail a secure access to a resource. It is managed with the client's knowledge and normally requires his/her cooperation. As an example, a person's access to a bank account at an automatic teller machine (ATM) may be asked to verify his fingerprint or look at a camera for face verification or speak into a microphone for voice authentication. Another example is the fingerprint readers of most modern laptop computers that allow access to the system only after fingerprint verification.

Person identification systems are more likely to operate covertly without the knowledge of the client. This can be

used, for example, to identify speakers in a recorded group conversation, or a criminal's fingerprint or voice is cross checked against a database of voices and fingerprints looking for a match.

Recognition systems have typically two phases: enrollment and test. During the enrollment phase, the client deliberately registers on the system one or more biometric traits. The system derives a number of features for these traits to form a client print, template, or model. During the test phase, whether identification or verification, the client is biometrically matched against the model(s).

This paper is solely concerned with the identity verification task. Thus, the two terms verification and recognition referred to herein are used interchangeably to indicate verification.

3.2. Biometric Modalities. Identity verification systems rely on multiple biometric modalities to match clients. These modalities include voice, facial geometry, fingerprint, signature, iris, retina, and hand geometry. Each one of these modalities has been extensively researched in literature. This paper focuses on the voice and the face modalities.

It has been established that multimodal identity verification systems outperform verification systems that rely on a single biometric modality [10, 11]. Such performance gain is more apparent in noisy environments; identity verification systems that rely solely on speech are affected greatly by the microphone type, the level of background noise (street noise, cafeteria atmosphere, ...), and the physical state of the speaker (sickness, mental state, ...). Identity verification systems based on the face modality is dependent on the video camera quality, the face brightness, and the physical appearance of the subject (hair style, beard, makeup, ...).

3.2.1. Voice. Voice verification, also known as speaker recognition, is a biometric modality that relies on features influenced by both the structure of a person's vocal tract and the speech behavioral characteristics. The voice is a widely acceptable modality for person verification and has been a subject for research for decades. There are two forms of speaker verification: text dependent (constrained mode), and text independent (unconstrained mode). Speaker verification is treated in Section 3.3.

3.2.2. Face. The face modality is a widely acceptable modality for person recognition and has been extensively researched. The face recognition process has matured into a science of sophisticated mathematical representations and matching processes. There are two predominant approaches to the face recognition problem: holistic methods and feature-based techniques. Face verification is described in Section 3.4.

3.3. Speaker Verification. The speech signal is an important biometric modality used in the audiovisual verification system. To process this signal a feature extraction module calculates relevant feature vectors from the speech waveform. On a signal window that is shifted at a regular rate a feature vector is calculated. Generally, cepstral-based feature



FIGURE 1: Screenshots from the BANCA database for two subjects in all three scenarios: Controlled (left), degraded (middle), and adverse (right).

vectors are used. A stochastic model is then applied to represent the feature vectors from a given speaker. To verify a claimed identity, new utterance feature vectors are generally matched against the claimed speaker model and against a general model of speech that may be uttered by any speaker, called the world model. The most likely model identifies if the claimed speaker has uttered the signal or not. In text independent speaker verification, the model should not reflect a specific speech structure, that is, a specific sequence of words. State-of-the art systems use Gaussian Mixture Models (GMMs) as stochastic models in text-independent mode. A tutorial on speaker verification is provided in [12].

3.3.1. Feature Extraction. The first part of the speaker verification process is the speech signal analysis. Speech is inherently a nonstationary signal. Consequently, speech analysis is normally performed on short fragments of speech where the signal is presumed stationary. To compensate for the signal truncation, a weighting signal is applied on each window.

Coding the truncated speech windows is achieved through variable resolution spectral analysis [13]. The most common technique employed is filter-bank analysis; it is a conventional spectral analysis technique that represents the signal spectrum with the log-energies using a filter-bank of overlapping band-pass filters.

The next step is cepstral analysis. The cepstrum is the inverse Fourier transform of the logarithm of the Fourier transform of the signal. A determined number of mel frequency cepstral coefficients (MFCCs) are used to represent the spectral envelope of the speech signal. They are derived from the filter-bank energies. To reduce the effects of signals recorded in different conditions, Cepstral mean subtraction and feature variance normalization is used. First- and second-order derivatives of extracted features are

appended to the feature vectors to account for the dynamic nature of speech.

3.3.2. Silence Detection. It is well known that the silence part of the signal alters largely the performance of a speaker verification system. Actually, silence does not carry any useful information about the speaker, and its presence introduces a bias in the score calculated, which deteriorates the system performance. Therefore, most of the speaker recognition systems remove the silence parts from the signal before starting the recognition process. Several techniques have been used successfully for silence removal. In our experiments, we suppose that the energy in the signal is a random process that follows a bi-Gaussian model, a first Gaussian modeling the energy of the silence part and the other modeling the energy of the speech part. Given an utterance and more specifically the computed energy coefficients, the bi-Gaussian model parameters are estimated using the EM algorithm. Then, the signal is divided into speech parts and silence parts based on a maximum likelihood criterion. Treatment of silence detection can be found in [14, 15].

3.3.3. Speaker Classification and Modeling. Each speaker possesses a unique vocal signature that provides him with a distinct identity. The purpose of speaker classification is to exploit such distinctions in order to verify the identity of a speaker. Such classification is accomplished by modeling speakers using a Gaussian Mixture Model (GMM).

Gaussian Mixture Models. A mixture of Gaussians is a weighted sum of M Gaussian densities

$$P(x|\lambda) = \sum_{i=1:M} \alpha_i f_i(x), \quad (3)$$

where x is an MFCC vector, $f_i(x)$ is a Gaussian density function, and α_i is the corresponding weights. Each Gaussian

is characterized by its mean μ_i and a covariance matrix Σ_i . A speaker model λ is characterized by the set of parameters $(\alpha_i, \mu_i, \Sigma_i)_{i=1:M}$.

For each client, two GMMs are used: the first corresponds to the distribution of the training set of speech feature vectors of that client, and the second represents the distribution of the training vectors of a defined “world model.”

To formulate the classification concept, assume that a speaker is presented along with an identity claim C . The feature vectors $V = \{\vec{v}_i\}_{i=1}^N$ are extracted. The average log likelihood of the speaker having identity C is calculated as

$$\mathcal{L}(X | \lambda_c) = \frac{1}{N} \sum_{i=1}^N \log p(\vec{x}_i | \lambda_c), \quad (4)$$

where $p(\vec{x}_i | \lambda_c) = \sum_{j=1}^{N_G} m_j \mathcal{N}(\vec{x}; \vec{\mu}_j, \Sigma_j)$, $\lambda = \{m_j \vec{\mu}_j, \Sigma_j\}_{j=1}^{N_G}$, and $\mathcal{N}(\vec{x}; \vec{\mu}_j, \Sigma_j) = (1/(2\pi)^{D/2} |\Sigma_j|^{1/2}) e^{-(1/2)(\vec{x} - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j)}$ is a multivariate Gaussian function with mean $\vec{\mu}_j$ and diagonal covariance matrix Σ_j , and D is the dimension of the feature space, λ_c is the parameter set for person C , N_G is the number of Gaussians, m_j = weight for Gaussian j , and $\sum_{j=1}^{N_G} m_j = 1$, $m_j \geq 0 \forall j$.

With a world model of w persons, the average log likelihood of a speaker being an impostor is found as

$$\mathcal{L}(X | \lambda_w) = \frac{1}{N} \sum_{i=1}^N \log p(\vec{x}_i | \lambda_w). \quad (5)$$

An opinion on the claim is then found: $\mathcal{O}(X) = \log \mathcal{L}(X | \lambda_c) - \log \mathcal{L}(X | \lambda_w)$.

As a final decision to whether the face belongs to the claimed identity, and given a certain threshold t , the claim is accepted when $\mathcal{O}(X) \geq t$ and rejected when $\mathcal{O}(X) < t$.

To estimate the GMM parameters λ of each speaker, the world model is adapted using a Maximum a Posteriori (MAP) adaptation [16]. The world model parameters are estimated using the Expectation Maximization (EM) algorithm [17].

GMM client training and testing is performed on the speaker verification toolkit BECARs [18]. BECARs implements GMMs with several adaptation techniques, for example, Bayesian adaptation, MAP, maximum likelihood linear regression (MLLR), and the unified adaptation technique defined in [19].

3.4. Face Verification. Face verification is a biometric person recognition technique used to verify (confirm or deny) a claimed identity based on a face image or a set of faces (or a video sequence). The process of automatic face recognition can be thought of as being comprised of four stages:

- (i) face detection, localization and segmentation;
- (ii) normalization;
- (iii) facial Feature extraction and tracking;
- (iv) classification (identification and/or verification).

These subtasks have been independently researched and surveyed in literature and are briefed next.

3.4.1. Face Detection. Face detection is an essential part of any face recognition technique. Given an image, face detection algorithms try to answer the following questions.

- (i) Is there a face in the image?
- (ii) If there is a face in the image, where is it located?
- (iii) What are the size and the orientation of the face?

Face detection techniques are surveyed in [20, 21].

The face detection algorithm used in this work has been introduced initially by Viola and Jones [22] and later developed further by Lienhart and Maydt [23]. It is a machine learning approach based on a boosted cascade of simple and rotated haar-like features for visual object detection.

3.4.2. Face Tracking in a Video Sequence. Face tracking in a video sequence is a direct extension of still image face detection techniques. However, the coherent use of both spatial and temporal information of faces makes the detection techniques more unique.

The technique used in this work employs the algorithm developed by Lienhart on every frame in the video sequence. However, three types of tracking errors are identified in a talking face video.

- (i) More than one face is detected, but only one actually exists in a frame.
- (ii) A wrong object is detected as a face.
- (iii) No faces are detected.

Figure 2 shows an example detection from the BANCA database [1], where two faces have been detected, one for the actual talking-face subject, and a false alarm.

The correction of these errors is done through the exploitation of spatial and temporal information in the video sequence as the face detection algorithm is run on every subsequent frame. The correction algorithm is summarized as follows.

- (a) *More than one face area detected.* The intersections of these areas with the area of the face of the previous frame are calculated. The area that corresponds to the largest calculated intersection is assigned as the face of the current frame. If the video frame in question is the first one in the video sequence, then the decision to select the proper face for that frame is delayed until a single face is detected at a later frame and verified with a series of subsequent face detections.
- (b) *No faces detected.* The face area of the previous frame is assigned as the face of the current frame. If the video frame in question is the first one in the video sequence, then the decision is delayed as explained in part (a).
- (c) *A wrong object detected as a face.* The intersection area with the previous frame face area is calculated. If this intersection ratio to the area of the previous face is less than a certain threshold, for example, 80%, the previous face is assigned as the face of the current frame.

3.4.3. Face Normalization. Normalizing face images is a required preprocessing step that aims at reducing the variability of different aspects in the face image such as contrast and illumination, scale, translation, rotation, and face masking. Many works in literature [24–26] have normalized face images with respect to translation, scale, and in-plane rotation, while others [27, 28] have also included masking and affine warping to properly align the faces. Craw and Cameron in [28] have used manually annotated points around shapes to warp the images to the mean shape, leading to shape-free representation of images useful in PCA classification.

The preprocessing stage in this work includes four steps.

- (i) Scaling the face image to a predetermined size (w_f , h_f).
- (ii) Cropping the face image to an inner-face, thus disregarding any background visual data.
- (iii) Disregarding color information by converting the face image to grayscale.
- (iv) Histogram equalization of the face image to compensate for illumination changes.

Figure 3 shows an example of the four steps.

3.4.4. Feature Extraction. The facial feature extraction technique used in this work uses DCT-*mod2* proposed by Sandereson and Paliwal in [29]. This technique is used in this work for its simplicity and performance in terms of computational speed and robustness to illumination changes.

A face image is divided into overlapping $N \times N$ blocks. Each block is decomposed in terms of orthogonal 2D DCT basis functions and is represented by an ordered vector of DCT coefficients:

$$\left[c_0^{(b,a)} c_1^{(b,a)} \dots c_{M-1}^{(b,a)} \right]^T, \quad (6)$$

where (b, a) represent the location of the block, and M is the number of the most significant retained coefficients. To minimize the effects of illumination changes, horizontal and vertical delta coefficients for blocks at (b, a) are defined as first-order orthogonal polynomial coefficients, as described in [29].

The first three coefficients c_0 , c_1 , and c_2 are replaced in (6) by their corresponding deltas to form a feature vector of size $M + 3$ for a block at (b, a) :

$$\left[\Delta^h c_0 \Delta^v c_0 \Delta^h c_1 \Delta^v c_1 \Delta^h c_2 \Delta^v c_2 c_3 c_4 \dots c_{M-1} \right]^T. \quad (7)$$

3.4.5. Face Classification. Face verification can be seen as a two-class classification problem. The first class is the case when a given face corresponds to the claimed identity (client), and the second is the case when a face belongs to an impostor. In a similar way to speaker verification, a GMM is used to model the distribution of face feature vectors for each person.

3.5. Fusion. It has been shown that biometric verification systems that combine different modalities outperform single modality systems [30]. A final decision on the claimed identity of a person relies on both the speech-based and the face-based verification systems. To combine both modalities, a fusion scheme is needed.

Various fusion techniques have been proposed and investigated in literature. Ben-Yacoub et al. [10] evaluated different binary classification approaches for data fusion, namely, Support Vector Machine (SVM), minimum cost Bayesian classifier, Fisher's linear discriminant analysis, C4.5 decision classifier, and multilayer perceptron (MLP) classifier. The use of these techniques is motivated by the fact that biometric verification is merely a binary classification problem. An overview of fusion techniques for audio-visual identity verification is provided in [31].

Other fusion techniques used include the weighted sum rule and the weighted product rule. It has been shown that the sum rule and support vector machines are superior when compared to other fusion schemes [10, 32, 33].

The weighted sum rule fusion technique is used in this study. The sum rule computes the audiovisual score s by weight averaging: $s = w_s s_s + w_f s_f$, where w_s and w_f are speech and face score weights computed so as to optimize the equal error rate (EER) on the training set. The speech and face scores must be in the same range (e.g., $\mu = 0$, $\sigma = 1$) for the fusion to be meaningful. This is achieved by normalizing the scores as follows:

$$s_{\text{norm}(s)} = \frac{s_s - \mu_s}{\sigma_s}, \quad s_{\text{norm}(f)} = \frac{s_f - \mu_f}{\sigma_f}. \quad (8)$$

4. Audiovisual Imposture

Audiovisual imposture is the deliberate modification of both speech and face of a person so as to make him sound and look like someone else. The goal of such an effort is to analyze the robustness of biometric identity verification systems to forgery attacks. An attempt is made to increase the acceptance rate of an impostor. Transformations of both modalities are treated separately below.

4.1. Speaker Transformation. Speaker transformation, also referred to as voice transformation, voice conversion, or speaker forgery, is the process of altering an utterance from a speaker (impostor) to make it sound as if it was articulated by a target speaker (client). Such transformation can be effectively used to replace the client's voice in a video to impersonate that client and break an identity verification system.

Speaker transformation techniques might involve modifications of different aspects of the speech signal that carries the speaker's identity such as the formant spectra, that is, the coarse spectral structure associated with the different phones in the speech signal [34], the excitation function, that is, the "fine" spectral detail [35], the prosodic features, that is, aspects of the speech that occur over timescales larger than individual phonemes, and the mannerisms such as particular word choice or preferred phrases, or all kinds

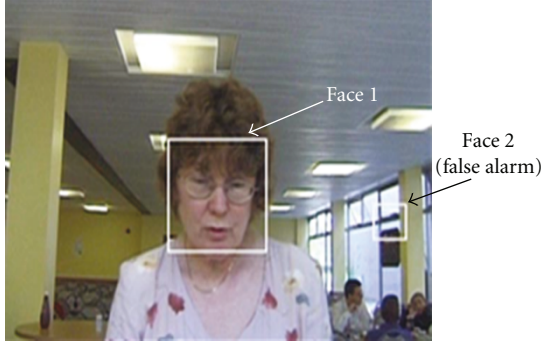


FIGURE 2: Face detection and tracking.

of other high-level behavioral characteristics. The formant structure and the vocal tract are represented by the overall spectral envelope shape of the signal, and thus they are major features to be considered in voice transformation [36].

Several voice transformation techniques have been proposed in literature. These techniques can be classified as text-dependent methods and text independent methods. In text-dependent methods, training procedures are based on parallel corpora, that is, training data have the source and the target speakers uttering the same text. Such methods include vector quantization [37, 38], linear transformation [36, 39], formant transformation [40], probabilistic transformation [41], vocal tract length normalization (VTLN) [42], and prosodic transformation [38]. In text-independent voice conversion techniques, the system trains on source and target speakers uttering different text. Techniques include text-independent VTLN [42], maximum likelihood constrained adaptation [43], and client memory indexation [44, 45].

The analysis part of a voice conversion algorithm focuses on the extraction of the speaker's identity. Next, a transformation function is estimated. At last, a synthesis step is achieved to replace the source speaker characteristics by those of the target speaker.

Consider a sequence of spectral vectors uttered by the source speaker (impostor) $X_s = [x_1, x_2, \dots, x_n]$, and a sequence pronounced by the target speaker comprising the same words $Y_t = [y_1, y_2, \dots, y_n]$. Voice conversion is based on the estimation of a conversion function \mathcal{F} that minimizes the mean square error $\epsilon_{\text{mse}} = E[\|y - \mathcal{F}(x)\|^2]$, where E is the expectation.

Two steps are useful to build a conversion system: training and conversion. In the training phase, speech samples from the source and the target speakers are analyzed to extract the main features. These features are then time aligned, and a conversion function is estimated to map the source and the target characteristics (Figure 4).

The aim of the conversion is to apply the estimated transformation rule to an original speech pronounced by the source speaker. The new utterance sounds like the same speech pronounced by the target speaker, that is, pronounced by replacing the source characteristics by those of the target voice. The last step is the resynthesis of the signal to reconstruct the source speech voice (Figure 5).

Voice conversion can be effectively used by an impostor to impersonate a target person and hide his identity in an attempt to increase the acceptance rate of the impostor by the identity verification system.

In this paper, *MixTrans*, a new mixture-structured bias voice transformation, is proposed and is described next.

4.1.1. MixTrans. A linear time-invariant transformation in the temporal domain is equivalent to a bias in the cepstral domain. However, speaker transformation may not be seen as a simple linear time-invariant transformation. It is more accurate to consider the speaker transformation as several linear time-invariant filters, each of them operating in a part of the acoustical space. This leads to the following form for the transformation:

$$\mathcal{T}_\theta(\mathbf{X}) = \sum_k \prod_k (\mathbf{X} + \mathbf{b}_k) = \sum_k \prod_k \mathbf{X} + \sum_k \prod_k \mathbf{b}_k = \mathbf{X} + \sum_k \prod_k \mathbf{b}_k, \quad (9)$$

where \mathbf{b}_k represents the k th bias, and \prod_k is the probability of being in the k th part of the acoustical space given the observation vector \mathbf{X} . \prod_k is calculated using a universal GMM modeling the acoustic space.

Once the transformation is defined, its parameters have to be estimated. We suppose that speech samples are available for both the source and the target speakers but do not correspond to the same text. Let λ be the stochastic model for a target client. λ is a GMM of the client. Let \mathbf{X} represent the sequence of observation vectors for an impostor (a source client). Our aim is to define a transformation $\mathcal{T}_\theta(\mathbf{X})$ that makes the source client vector resemble the target client. In other words, we would like to have the source vectors be best represented by the target client model λ through the application of the transformation $\mathcal{T}_\theta(\mathbf{X})$. In this context the Maximum likelihood criterion is used to estimate the transformation parameters:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\mathcal{T}_\theta(\mathbf{X}) | \lambda). \quad (10)$$

Since λ is a GMM, $\mathcal{T}_\theta(\mathbf{X})$ is a transform of the source vectors \mathbf{X} , and $\mathcal{T}_\theta(\mathbf{X})$ depends on another model λ_w , then $\mathcal{L}(\mathcal{T}_\theta(\mathbf{X}) | \lambda)$ in (10) can be written as

$$\begin{aligned} \mathcal{L}(\mathcal{T}_\theta(\mathbf{X}) | \lambda) &= \prod_{t=1}^T \mathcal{L}(\mathcal{T}_\theta(\mathbf{X}_t) | \lambda) \\ &= \prod_{t=1}^T \sum_{m=1}^M \frac{1}{(2\pi)^{D/2} |\Sigma_m|^{1/2}} e^{-(1/2)(\mathcal{T}_\theta(\mathbf{X}_t) - \mu_m)^T \Sigma_m^{-1} (\mathcal{T}_\theta(\mathbf{X}_t) - \mu_m)} \\ &= \prod_{t=1}^T \sum_{m=1}^M \frac{1}{(2\pi)^{D/2} |\Sigma_m|^{1/2}} \\ &\quad \times e^{-(1/2)(\mathbf{X}_t + \sum_{k=1}^K \prod_{ki} b_k - \mu_m)^T \Sigma_m^{-1} (\mathbf{X}_t + \sum_{k=1}^K \prod_{ki} b_k - \mu_m)}. \end{aligned} \quad (11)$$

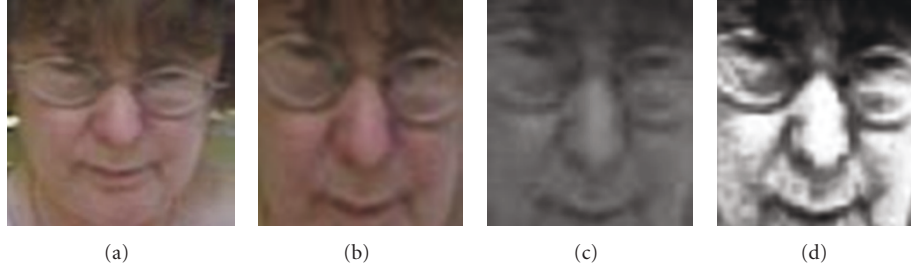


FIGURE 3: Preprocessing face images. (a) Detected face. (b) Cropped face (inner face). (c) Grayscale face. (d) Histogram-equalized face.

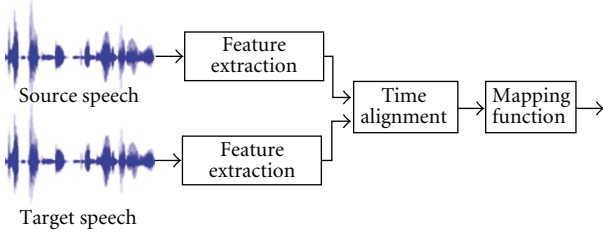


FIGURE 4: Training.

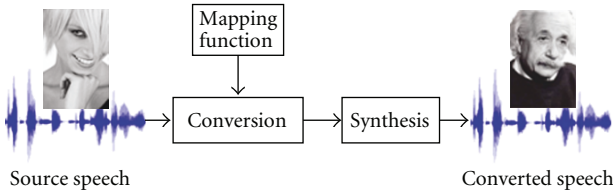


FIGURE 5: Conversion.

Finding $\{b_k\}$ such that (11) is maximized is found through the use of the EM algorithm. In the expectation “E” step, the probability α_{mt} of component m is calculated. Then, at the maximization “M” step, the log-likelihood is optimized dimension by dimension for a GMM with a diagonal covariance matrix:

$$ll = \sum_{t=1}^T \sum_{m=1}^M \alpha_{mt} \left[\log \frac{1}{\sigma_m \sqrt{2\pi}} - \frac{1}{2} \frac{(\mathbf{X}_t + \sum_{k=1}^K \Pi_{kt} \mathbf{b}_k - \mu_m)^2}{\sigma_m^2} \right]. \quad (12)$$

Maximizing

$$\frac{\partial ll}{\partial b_l} = 0 \Rightarrow - \sum_{t=1}^T \sum_{m=1}^M \alpha_{mt} \frac{(\mathbf{X}_t + \sum_{k=1}^K \Pi_{kt} \mathbf{b}_k - \mu_m) \Pi_{lt}}{\sigma_m^2} = 0, \quad \text{for } l = 1 \dots K, \quad (13)$$

then,

$$\sum_{t=1}^T \sum_{m=1}^M \frac{\alpha_{mt} \Pi_{lt}}{\sigma_m^2} (\mathbf{X}_t - \mu_m) = - \sum_{t=1}^T \sum_{m=1}^M \sum_{k=1}^K \frac{\alpha_{mt} \Pi_{kt} \Pi_{lt} \mathbf{b}_k}{\sigma_m^2}, \quad \text{for } l = 1 \dots K,$$

$$\sum_{t=1}^T \sum_{m=1}^M \frac{\alpha_{mt} \Pi_{lt}}{\sigma_m^2} (\mathbf{X}_t - \mu_m) = - \sum_{k=1}^K \mathbf{b}_k \sum_{m=1}^M \sum_{t=1}^T \frac{\alpha_{mt} \Pi_{lt} \Pi_{kt}}{\sigma_m^2}, \quad \text{for } l = 1 \dots K, \quad (14)$$

and finally, in matrix notation,

$$- \left(\sum_m \sum_t \frac{\alpha_{mt} \Pi_{lt} \Pi_{kt}}{\sigma_m^2} \right) (\mathbf{b}_k) = \left(\sum_m \sum_t \frac{\alpha_{mt} \Pi_{lt} (\mathbf{X}_t - \mu_m)}{\sigma_m^2} \right). \quad (15)$$

This matrix equation is solved at every iteration of the EM algorithm.

4.1.2. Speech Signal Reconstruction. It is known that the cepstral domain is appropriate for classification due to the physical significance of the Euclidean distance in this space [13]. However, the extraction of cepstral coefficients from the temporal signal is a nonlinear process, and the inversion of this process is not uniquely defined. Therefore, a solution has to be found in order to take the advantage of the good characteristics of the cepstral space while applying the transformation in the temporal domain.

Several techniques have been proposed to overcome this problem. In [46], harmonic plus noise analysis has been used for this purpose. Instead of trying to find a transformation allowing the passage from the cepstral domain to the temporal domain, a different strategy is adopted. Suppose that an intermediate space exists where transformation could be directly transposed to the temporal domain. Figure 6 shows the process where the temporal signal goes through a two-step feature extraction process leading to the cepstral coefficients that may be easily transformed into target speaker-like cepstral coefficients by applying the transformation function $\mathcal{T}_\theta(\mathbf{X})$ as discussed previously.

The transformation trained on the cepstral domain cannot be directly projected to the temporal domain since the feature extraction module $(\mathcal{F}_2 \circ \mathcal{F}_1)$ is highly nonlinear.

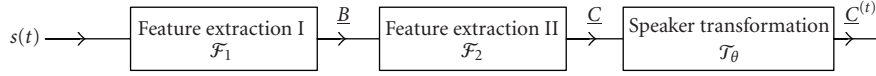


FIGURE 6: Steps from signal to transformed cepstral coefficients.

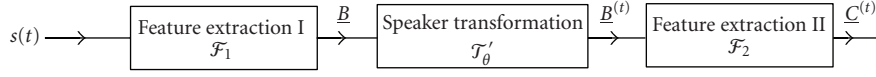


FIGURE 7: Steps from signal to transformed cepstral coefficients when transformation is applied in a signal-equivalent space.

However, a speaker transformation determined in the \underline{B} space may be directly projected in the signal space, for example, \underline{B} space may be the spectral domain. But, for physical significance it is better to train the transformation in the cepstral domain. Therefore, we suppose that another transformation $\mathcal{T}'_{\theta}(\mathbf{X})$ exists in the \underline{B} space leading to the same transformation in the cepstral domain satisfying thereby the two objectives: transformation of the signal and distance measurement in the cepstral domain. This is shown in Figure 7.

This being defined, the remaining issue is how to estimate the parameters θ of the transformation $\mathcal{T}'_{\theta}(\mathbf{X})$ in order to get the same transformation result as in the cepstral domain. This is detailed next.

4.1.3. Estimating Signal Transformation Equivalent to a Calculated Cepstral Transformation. The transformation in the cepstral domain is presumably determined; the idea is to establish a transformation in the \underline{B} space leading to cepstral coefficients similar to the one resulting from the cepstral transformation.

Let $\hat{\underline{C}}^{(t)}$ represent the cepstral vector obtained after the application of the transformation in the \underline{B} domain, and let $\underline{C}^{(t)}$ represent the cepstral vector obtained when applying the transformation in the cepstral domain. The difference defines an error vector:

$$\underline{e} = \underline{C}^{(t)} - \hat{\underline{C}}^{(t)}. \quad (16)$$

The quadratic error can be written as

$$E = |\underline{e}|^2 = \underline{e}^T \underline{e}. \quad (17)$$

Starting from a set of parameters for \mathcal{T}'_{θ} , the gradient algorithm may be applied in order to minimize the quadratic error E . For every iteration of the algorithm the parameter θ is updated using

$$\theta^{(i+1)} = \theta^{(i)} - \mu \frac{\partial E}{\partial \theta}, \quad (18)$$

where μ is the gradient step.

The gradient of the error with respect to parameter θ is given by

$$\frac{\partial E}{\partial \theta} = -2\underline{e}^T \frac{\partial \hat{\underline{C}}^{(t)}}{\partial \theta}. \quad (19)$$

Finally, the derivative of the estimated transformed cepstral coefficient with respect to θ can be obtained using a gradient descent

$$\frac{\partial \hat{\underline{C}}^{(t)}}{\partial \theta} = \frac{\partial \hat{\underline{C}}^{(t)T}}{\partial \underline{B}^{(t)}} \frac{\partial \underline{B}^{(t)}}{\partial \theta}. \quad (20)$$

In order to illustrate this principle, let us consider the case of MFCC analysis leading to the cepstral coefficients. In this case, \mathcal{F}_1 is just the Fast Fourier Transform (FFT) followed by the power spectral calculation (the phase being kept constant). \mathcal{F}_2 is the filterbank integration in the logarithm scale followed by the inverse DCT transform. We can write

$$\hat{C}_l^{(t)} = \sum_{k=1}^K \log \left(\sum_{i=1}^N a_i^{(k)} B_i^{(k)} \right) \cos \left(2\pi l \frac{f_k}{F} \right), \quad (21)$$

$$B_i^{(t)} = B_i \cdot \theta_i,$$

where $\{a_i\}$ are the filter-bank coefficients, f_k the central frequencies of the filter-bank, and θ_i is the transfer function at frequency bin i of the transformation $\mathcal{T}'_{\theta}(\mathbf{X})$.

Using (21), it is straightforward to compute the derivatives in (20):

$$\frac{\partial \hat{C}_l^{(t)}}{\partial B_j^{(t)}} = \sum_{k=1}^K \frac{a_j^{(k)}}{\sum_{i=1}^N a_i^{(k)} B_i^{(t)}} \cos \left(2\pi l \frac{f_k}{F} \right), \quad (22)$$

$$\frac{\partial B_i^{(t)}}{\partial \theta_j} = B_j \delta_{ij}.$$

Equations (19), (20), and (22) allow the implementation of this algorithm in the case of MFCC.

Once $\mathcal{T}'_{\theta}(\mathbf{X})$ completely defined, the transformed signal may be determined by applying an inverse FFT to $B(t)$ and using the original phase to recompose the signal window. In order to consider the overlapping between adjacent windows, the Overlap and Add (OLA) algorithm is used [47].

4.1.4. Initializing the Gradient Algorithm. The previous approach is computationally expensive. Actually, for each signal window, that is, from 10 milliseconds to 16 milliseconds, a gradient algorithm is to be applied. In order to alleviate this high computational algorithm, a solution consists in finding a good initialization of the gradient algorithm. This may be obtained by using an initial value for the transformation $\mathcal{T}'_{\theta}(\mathbf{X})$, the transformation obtained for the previous signal window.

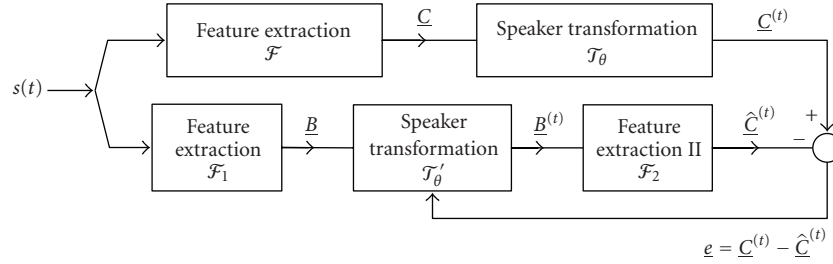


FIGURE 8: Signal-level transformation parameters tuned with a gradient descent algorithm.

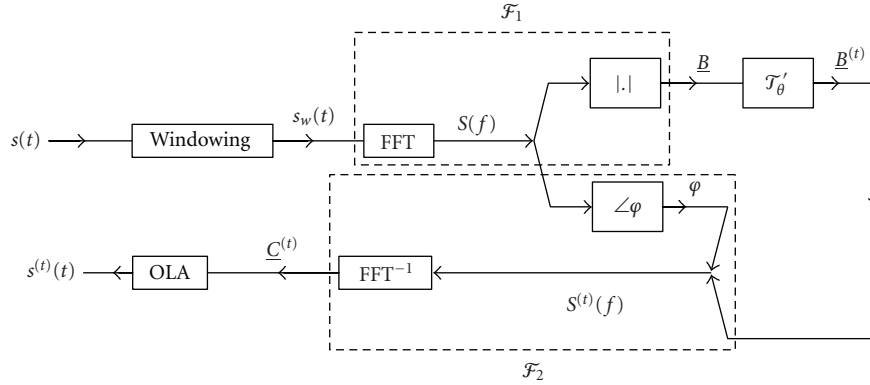


FIGURE 9: Speech signal feature extraction, transformation, and reconstruction.

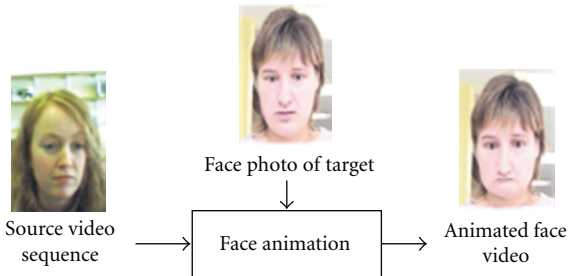


FIGURE 10: Face animation.

4.2. Face Animation. To complete the scenario of audiovisual imposture, speaker transformation is coupled with face transformation. It is meant to produce synthetically an “animated” face of a target person, given a still photo of his face and some animation parameters defined by a source video sequence. Figure 10 depicts the concept.

The face animation technique used in this paper is MPEG-4 compliant, which uses a very simple thin-plane spline warping function defined by a set of reference points on the target image, driven by a set of corresponding points on the source image face. This technique is described next.

4.2.1. MPEG-4 2D Face Animation. MPEG-4 is an object-based multimedia compression standard, which defines a standard for face animation [48]. It specifies 84 feature points (Figure 11) that are used as references for Facial Animation Parameters (FAPs). 68 FAPs allow the representation of facial expressions and actions such as head motion and mouth and

eye movements. Two FAP groups are defined, visemes (FAP group 1) and expressions (FAP group 2). Visemes (FAP1) are visually associated with phonemes of speech; expressions (FAP2) are joy, sadness, anger, fear, disgust, and surprise.

An MPEG-4 compliant system decodes an FAP stream and animates a face model that has all feature points properly determined. In this paper, the animation of the feature points is accomplished using a simple thin-plate spline warping technique and is briefly described next.

4.2.2. Thin-Plate Spline Warping. The thin-plate spline (TPS), initially introduced by Duchon [49], is a geometric mathematical formulation that can be applied to the problem of 2D coordinate transformation. The name *thin-plate spline* indicates a physical analogy to bending a thin sheet of metal in the vertical z direction, thus displacing x and y coordinates on the horizontal plane.

Given a set of data points $\{w_i, i = 1, 2, \dots, K\}$ in a 2D plane—for our case, MPEG-4 facial feature points—a radial basis function is defined as a spatial mapping that maps a location x in space to a new location $f(x) = \sum_{i=1}^K c_i \phi(\|x - w_i\|)$, where $\{c_i\}$ is a set of mapping coefficients, and the kernel function $\phi(r) = r^2 \ln r$ is the thin-plate spline [50]. The mapping function $f(x)$ is fit between corresponding sets of points $\{x_i\}$ and $\{y_i\}$ by minimizing the “bending energy” I , defined as the sum of squares of the second derivatives:

$$I[f(x, y)] = \iint_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right] dx dy. \quad (23)$$

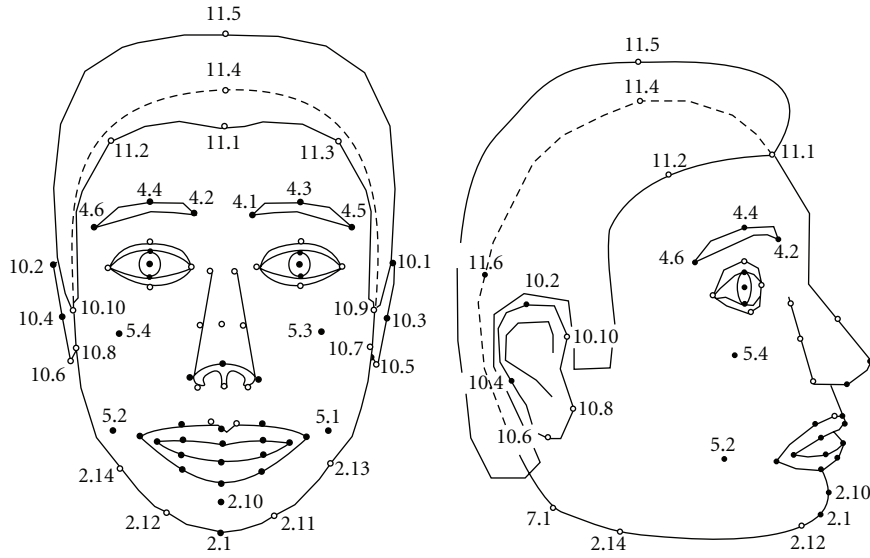


FIGURE 11: MPEG-4 feature points.

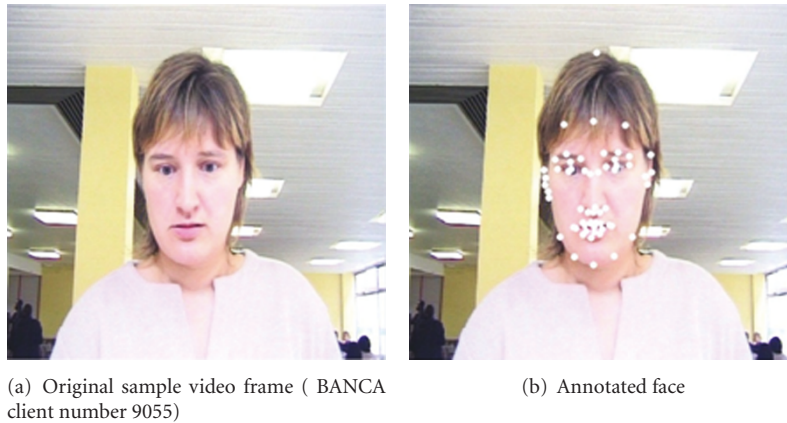


FIGURE 12: Feature point annotation on the BANCA database.

5. Effects of Imposture on Verification—Experimental Results

To test the robustness of IV systems, a state-of-the-art baseline audio-visual IV system is built. This system follows the BANCA “P” protocol and is based on a classical GMM approach for both speech and face modalities. It is completely independent from the voice and face transformations described above.

5.1. Verification Experiments

5.1.1. Speaker Verification. For speech, feature extraction and silence detection is first performed, as described in Sections 3.3.1 and 3.3.2. Then GMM speaker classification is performed with 256 Gaussians. The world model of BANCA is adapted using MAP adaptation, and its parameters estimated using the EM algorithm, as discussed in Section 3.3.3

above. The world model is used as a Universal Background Model (UBM) for training to amplify the variability between different speakers. In fact, to improve the performance of the IV system, we use a larger UBM by combining BANCA world model and g2 when training and testing is performed on g1 and vice versa. This is possible because g1 and g2 of BANCA are totally independent. Client models are then adapted from the UBM using speech features from the enrollment set of BANCA. To verify a claimed identity of a test speaker, his/her features are extracted and compared to both the UBM and the GMM of the client. The average log likelihood is calculated, and an acceptance or a rejection decision is taken as described in Section 3.3.3. A total of 234 true client tests and 312 impostor tests (per group) were performed in compliance with BANCA’s “P” protocol. Figure 14(a) shows the DET curves for speaker verification on g1 and g2, with an EER of 4.38% and 4.22%, respectively.



FIGURE 13: Selected frames from an animated face with various expressions.

5.1.2. Face Verification Experiments. Face verification is based on extracting facial features from a video sequence as described in Section 3.4. First, the face tracking module extracts faces in all frames and retains only 5 of them for training and/or testing. The 5 frames selected are equally distributed across the video sequence so as to have a good sample of faces. These faces are then resized to 48×64 , gray-scaled, cropped to 36×40 , and then histogram-equalized. Then DCT feature extraction follows. Neighboring blocks of 8×8 with an overlap of 50% is used. M , the number of retained coefficients, is fixed at 15 [29]. In a similar way to speaker verification, GMMs are used to model the distribution of face feature vectors for each person.

For the same BANCA “P” protocol, and a total of 234 true clients and 312 impostor tests (per group per frame—5 frames per video) the DET curves for face verification are shown in Figure 14(b) with an EER of 23.5% and 22.2% for g1 and g2, respectively.

5.1.3. Score Fusion. Figure 14(c) shows an improvement of the verification by score fusion of both modalities, with an EER of 4.22% for g1 and 3.47% for g2. The optimized weights w_s and w_f are integers 8 and 3, respectively, as described in Section 3.5.

5.2. Transformation Experiments. BANCA defines in its protocols imposture attempts during which a speaker proclaims in his/her own voice and face to be someone else. This “zero-effort” imposture is unrealistic, and any text-independent verification system should be able to detect easily the forgery by contrasting the impostor model against the claimed identity model. To make the verification more difficult, transformation of both voice and face is performed.

5.2.1. Voice Conversion Experiments. BANCA has total of 312 impostor attacks per group in which the speaker claims in his own words to be someone else. These attempts are replaced by the transformed voices as described in Section 4.1. For each attempt, MFCC analysis is performed, and transformation coefficients are calculated in the cepstral domain using the EM algorithm. Then the signal transformation parameters are estimated using a gradient descent algorithm. The transformed voice signal is then reconstructed with an

inverse FFT and OLA as described in Section 4.1.3. The pitch of the transformed voice had to be adjusted to match better the target speaker’s pitch. Verification experiments are repeated with the transformed voices. The result is an increase of the EER from 4.38% to 10.6% on g1 and from 4.22% to 12.1% on g2 (Figure 14(a)).

5.2.2. Face Conversion Experiments. Given a still picture of the face of a target person, the MPEG-4 facial feature points are first manually annotated as shown in Figure 12. A total of 61 feature points out of 83 specified by MPEG-4 are annotated, the majority of which belong to the eyes and the mouth regions. Others have less impact on FAPs or do not affect them at all.

The FAPs used in the experiments correspond to a subset of 33 out of the 68 FAPs defined by MPEG-4. Facial actions related to head movement, tongue, nose, ears, and jaws are not used. The FAPs used correspond to mouth, eye, and eyebrow movements, for example, horizontal displacement of right outer lip corner (`stretch_r_cornerlip_o`), vertical displacement of top right eyelid (`close_t_r_eyelid`), and vertical displacement of left outer eyebrow (`raise_l_o_eyebrow`). Figure 13 shows animated frames simulating the noted expressions.

A synthesized video sequence is generated by deforming a face from its neutral state according to determined FAP values at a rate of 25 frames per second. For the experiments presented in this work, these FAPs are selected so as to produce a realistic talking head that is not necessarily synchronized with the associated transformed speech. The only association with speech is the duration of the video sequence, which corresponds to the total time of speech. The detection and the measure of the level of audiovisual speech synchrony is not treated in this work but has been reported in [51–53] to improve the verification performance.

BANCA has total of 312 impostor attacks per group in which the speaker claims in his own words and facial expressions to be someone else. These are replaced by the synthetically animated videos with the transformed speech. The experiments have shown a deterioration of the performance from an EER from (23.5%, 22.2%) on (g1, g2) to (37.6%, 33.0%) (Figure 14(b)) for face, and from (4.22%, 3.47%) to (11.0%, 16.1%) for the audio-visual system (Figure 14(c)).

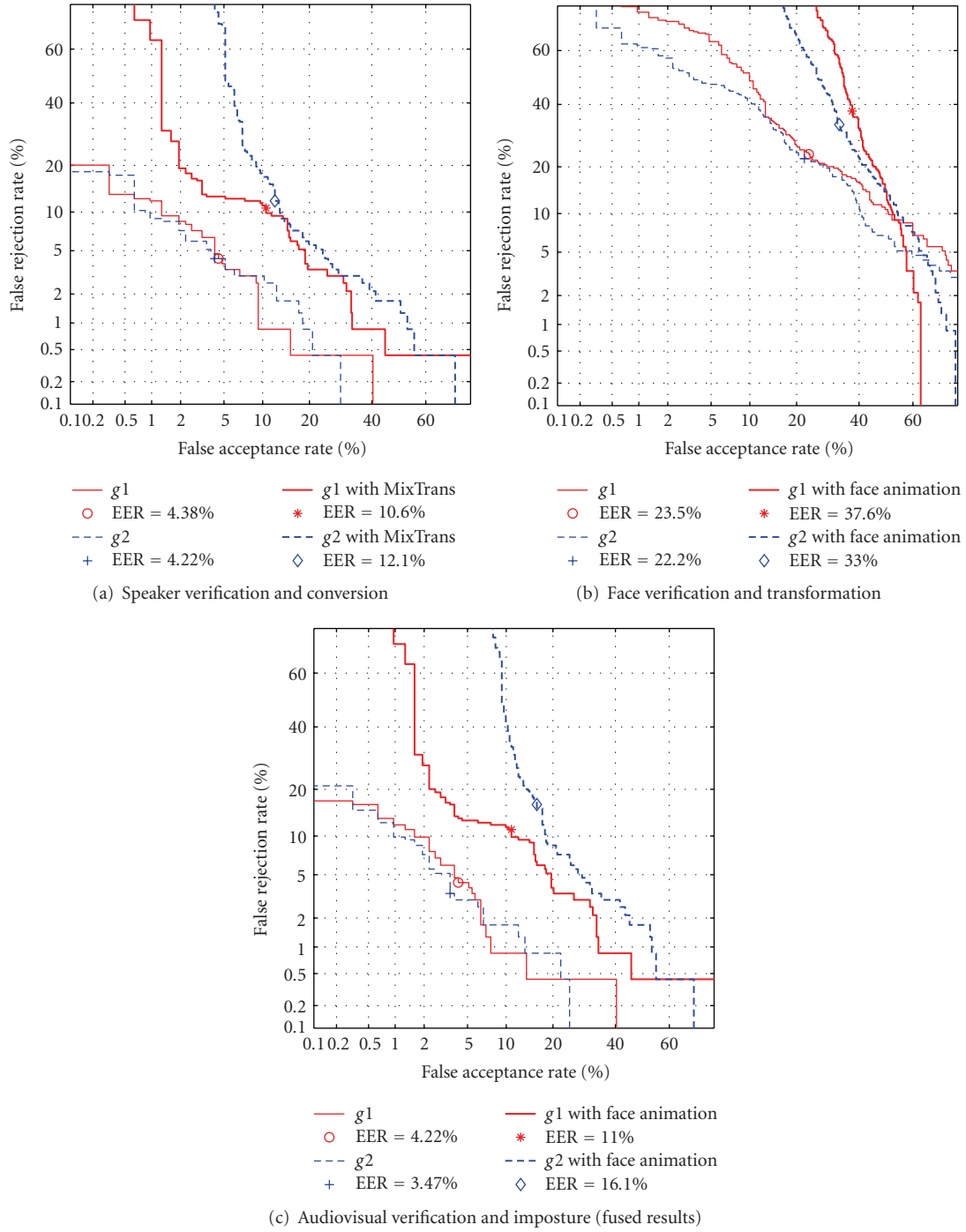


FIGURE 14: Audiovisual verification and imposture results on BANCA.

6. Conclusion

This paper provides a review of biometric identity verification techniques and describes their evaluation and robustness to imposture. It proposes *MixTrans*, a mixture-structured bias voice transformation technique in the cepstral domain, which allows a transformed audio signal to be estimated and reconstructed in the temporal domain. It also couples the audio conversion with an MPEG-4 compliant

face animation system that warps facial feature points using a simple thin-plate spline. The proposed audiovisual forgery is completely independent from the baseline audiovisual IV system and can be used to attack any other audiovisual IV system. The Results drawn from the experiments show that state-of-the-art verification systems are vulnerable to forgery, with an EER average increase from 3.8% to 13.5%. This increase clearly shows that such attacks represent a serious challenge and a security threat to audio-visual IV

systems. The results show that state-of-the-art IV systems are vulnerable to forgery attacks, which indicate more impostor acceptance, and, for the same threshold, more genuine client denial. This should drive more research towards more robust IV systems.

References

- [1] E. Bailly-Baillié, S. Bengio, F. Bimbot, et al., "The BANCA database and evaluation protocol," in *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 625–638, Guildford, UK, June 2003.
- [2] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "Xm2vtsdb: the extended m2vts database," in *Proceedings of the 2nd International Conference on Audio and Video-Based Biometric Person Authentication (AVBPA '99)*, pp. 72–77, Washington, DC, USA, March 1999.
- [3] J. S. D. Mason, F. Deravi, C. C. Chibelushi, and S. Gandon, "Project: david (digital audio visual integrated database)," Tech. Rep., Department of Electrical and Electronic Engineering, University of Wales Swansea, Swansea, UK, 1996, <http://eegalilee.swan.ac.uk>.
- [4] S. Garcia-Salicetti, C. Beumier, G. Chollet, et al., "Biomet: a multimodal person authentication database including face, voice, fingerprint, hand and signature modalities," in *Proceedings of the 4th IAPR International Conference on Audio and Video-Based Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 845–853, Guildford, UK, June 2003.
- [5] A. C. Morris, J. Koreman, H. Sellaheewa, et al., "The secure-phone pda database, experimental protocol and automatic test procedure for multimodal user authentication," Tech. Rep., The SecurePhone Project, 2006.
- [6] J. P. Egan, *Signal Detection Theory and ROC Analysis*, Series in Cognition and Perception, Academic Press, New York, NY, USA, 1975.
- [7] T. Fawcett, "Roc graphs: notes and practical considerations for researchers," 2004, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.9777>.
- [8] S. Bengio, J. Mariethoz, and M. Keller, "The expected performance curve," in *Proceedings of the 2nd International Conference on Machine Learning, ICML, Workshop on ROC Analysis in Machine Learning*, Bonn, Germany, August 2005.
- [9] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, vol. 4, pp. 1895–1898, Rhodes, Greece, September 1997.
- [10] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1065–1074, 1999.
- [11] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.
- [12] F. Bimbot, J.-F. Bonastre, C. Fredouille, et al., "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.
- [13] C. Chouzenoux, *Analyse spectrale à résolution variable. Application au signal de parole*, Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, 1982.
- [14] J. Trmal, J. Zelinka, J. Psutka, and L. Müller, "Comparison between GMM and decision graphs based silence/speech detection method," in *Proceedings of the 11th International Conference Speech and Computer (SPECOM '06)*, pp. 376–379, Anatalya, St. Petersburg, Russia, June 2006.
- [15] D. R. Paoletti and G. Erten, "Enhanced silence detection in variable rate coding systems using voice extraction," in *Proceedings of the 43rd IEEE Midwest Symposium on Circuits and Systems (MWSCAS '00)*, vol. 2, pp. 592–594, Lansing, Mich, USA, August 2000.
- [16] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
- [17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
- [18] R. Blouet, C. Mokbel, H. Mokbel, E. Sánchez Soto, G. Chollet, and H. Greige, "Becars: a free software for speaker verification," in *Proceedings of the Speaker and Language Recognition Workshop (ODYSSEY '04)*, pp. 145–148, Toledo, Spain, May-June 2004.
- [19] C. Mokbel, "Online adaptation of HMMs to real-life conditions: a unified framework," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 342–357, 2001.
- [20] E. Hjelmås and B. K. Low, "Face detection: a survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.
- [21] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.
- [22] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 1, pp. 511–518, Kauai, Hawaii, USA, December 2001.
- [23] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '02)*, vol. 1, pp. 900–903, Rochester, NY, USA, September 2002.
- [24] P. N. Belhumeur and D. J. Kriegman, "What is the set of images of an object under all possible lighting conditions?" *International Journal of Computer Vision*, vol. 28, no. 3, pp. 270–277, 1998.
- [25] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.
- [26] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [27] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.
- [28] I. Craw and P. Cameron, "Manifold caricatures: on the psychological consistency of computer face recognition," in *Proceedings of the British Machine Vision Conference (BMVC '92)*, pp. 498–507, Springer, Leeds, UK, September 1992.
- [29] C. Sanderson and K. K. Paliwal, "Fast feature extraction method for robust face verification," *Electronics Letters*, vol. 38, no. 25, pp. 1648–1650, 2002.
- [30] J. Kittler, "Combining classifiers: a theoretical framework," *Pattern Analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.

- [31] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing*, vol. 14, no. 5, pp. 449–480, 2004.
- [32] V. Chatzis, A. G. Bors, and I. Pitas, "Multimodal decision-level fusion for person authentication," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 29, no. 6, pp. 674–680, 1999.
- [33] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, and J. Gonzalez-Rodriguez, "A comparative evaluation of fusion strategies for multimodal biometric verification," in *Proceedings of the 4th IAPR International Conference on Audio and Video-Based Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 830–837, Guildford, UK, June 2003.
- [34] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 1, pp. 285–288, Seattle, Wash, USA, May 1998.
- [35] B. Yegnanarayana, K. Sharat Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 1, pp. 409–412, Salt Lake City, Utah, USA, May 2001.
- [36] A. Kain, *High resolution voice transformation*, Ph.D. thesis, OGI School of Science and Engineering, Oregon Health & Science University, Portland, Ore, USA, 2001.
- [37] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '88)*, vol. 1, pp. 655–658, New York, NY, USA, April 1988.
- [38] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Communication*, vol. 28, no. 3, pp. 211–226, 1999.
- [39] H. Ye and S. Young, "Perceptually weighted linear transformations for voice conversion," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH '03)*, vol. 4, pp. 2409–2412, Geneva, Switzerland, September 2003.
- [40] E. Turajlic, D. Rentzos, S. Vaseghi, and C.-H. Ho, "Evaluation of methods for parameteric formant transformation in voice conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 724–727, Hong Kong, April 2003.
- [41] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.
- [42] D. Sundermann, H. Ney, and H. Hoge, "Vtln-based cross-language voice conversion," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pp. 676–681, St. Thomas, Virgin Islands, USA, December 2003.
- [43] A. Mouchtaris, J. van der Spiegel, and P. Mueller, "Non-parallel training for voice conversion by maximum likelihood constrained adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 1, pp. 1–4, Montreal, Canada, May 2004.
- [44] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using ALISP: indexation in a client memory," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, pp. 17–20, Philadelphia, Pa, USA, March 2005.
- [45] H. Ye and S. Young, "Voice conversion for unknown speakers," in *Proceedings of the 8th International Conference of Spoken Language Processing (ICSLP '04)*, pp. 1161–1164, Jeju Island, South Korea, October 2004.
- [46] Y. Stylianou and O. Cappe, "A system for voice conversion based on probabilistic classification and a harmonic plus noise model," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 1, pp. 281–284, Seattle, Wash, USA, May 1998.
- [47] A. V. Oppenheim and R. W. Schaffer, *Digital Signal Processing*, Prentice-Hall, Upper Saddle River, NJ, USA, 1975.
- [48] A. M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Signal Processing: Image Communication*, vol. 15, no. 4-5, pp. 387–421, 2000.
- [49] J. Duchon, "Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces," *RAIRO: Analyse Numérique*, vol. 10, no. 12, pp. 5–12, 1976.
- [50] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.
- [51] H. Bredin and G. Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, vol. 2, pp. 233–236, Honolulu, Hawaii, USA, April 2007.
- [52] H. Bredin and G. Chollet, "Audiovisual speech synchrony measure: application to biometrics," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 70186, 11 pages, 2007.
- [53] H. Bredin and G. Chollet, "Making talking-face authentication robust to deliberate imposture," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '08)*, pp. 1693–1696, Las Vegas, Nev, USA, April 2008.